Table 1: List of 50 models used in the simulation study

	Distribution	Parameter of interest	Other parameters
1	Bernoulli(p)	p	
2	Discrete Uniform(N)	N	
3	Geometric(p)	p	
4	Negtive Binomial(r, p)	r	$p = 0.3$
5	Exponential(λ)	λ	
6	Normal(μ, σ)	μ	$\sigma = 1$
7	Poisson(λ)	λ	
8	Beta(α, β)	α	$\beta = 3$
9	Weibull(λ, k)	k	$\lambda = 3$
10	Double Exponential(μ, λ)	μ	$\lambda = 3$
11	Chi Square(k)	k	
12	F(d_1, d_2)	d_2	$d_1 = 3$
13	Gamma(α, β)	β	$\alpha = 0.5$
14	Logistic(μ, s)	μ	$s = 0.5$
15	Lognormal(μ, σ)	μ	$\sigma = 0.5$
16	Pareto(x_m, α)	x_m	$\alpha = 2$
17	Student's t(ν, ncp)	ν	$ncp = 2$
18	Uniform(a, $a + 2$)	a	
19	Hypergeometric(m, n, k)	n	$m = 3, k = 2$
20	Binomial(n, p)	n	$p = 0.5$
21	One-Inflated Logarithmic($shape$, $pstr_1$)	$shape$	$pstr_1 = 0$
22	Triangle(θ, $lower$, $upper$)	θ	$lower = 0$, $upper = 33$
23	Wilcoxon Signed Rank Statistic(n)	n	
24	Benini(y_0, $shape$)	$shape$	$y_0 = 1$
25	Beta-Geometric($shape_1$, $shape_2$)	$shape_2$	$shape_1 = 5$
26	Beta-Normal($shape_1$, $shape_2$, $mean$, sd)	$shape_1$	$shape_2 = 10$, $mean = 5$, $sd = 11$
27	Birnbaum-Saunders($scale$, $shape$)	$shape$	$scale = 1$
28	Dagum($scale$, $shape_{1.a}$, $shape_{2.p}$)	$shape_{1.a}$	$scale = 1$, $shape_{2.p} = 2$
29	Frechet($location$, $scale$, $shape$)	$shape$	$location = 0$, $scale = 1$
30	Dirichlet($shape_1$, $shape_2$, $shape_3$)	$shape_1$	$shape_2 = 2$, $shape_3 = 4$
31	Huber's Least Favourable(k, μ, σ)	k	$\mu = 0$, $\sigma = 1$
32	Gumbel($location$, $scale$)	$scale$	$location = 1$
33	Gompertz($scale$, $shape$)	$shape$	$scale = 1$
34	Kumaraswamy($shape_1$, $shape_2$)	$shape_2$	$shape_1 = 10$
35	Laplace($location$, $scale$)	$scale$	$location = 5$
36	Log-Gamma($location$, $scale$, $shape$)	$scale$	$location = 0$, $shape = 2$
37	Lindley(θ)	θ	
38	Lomax($scale$, $shape_{3.q}$)	$shape_{3.q}$	$scale$
39	Makeham($scale$, $shape$, ϵ)	$shape$	$scale = 0$, $\epsilon = 0$
40	Maxwell($rate$)	$rate$	
41	Nakagami($scale$, $shape$, $Smallno$)	$shape$	$scale = 1$, $Smallno = 1.0e - 6$
42	Perks($scale$, $shape$)	$shape$	$scale = 1$
43	Rayleigh($scale$)	$scale$	
44	Rice(σ, vee)	vee	$\sigma = 1$
45	Simplex(μ, $dispersion$)	$dispersion$	$\mu = 0.5$
46	Singh-Maddala($scale$, $shape_{1.a}$, $shape_{3.q}$)	$shape_3.q$	$scale = 1$, $shape_{1.a} = 5$
47	Skellam(μ_1, μ_2)	μ_2	$\mu_1 = 5$
48	Tobit($mean$, sd, $lower$, $upper$)	$mean$	$sd = 1$, $lower = 0$, $upper = Inf$
49	Paralogistic($scale$, $shape_{1.a}$)	$scale_{1.a}$	$scale = 1$
50	Zipf(N, $shape$)	$shape$	$N = 10$

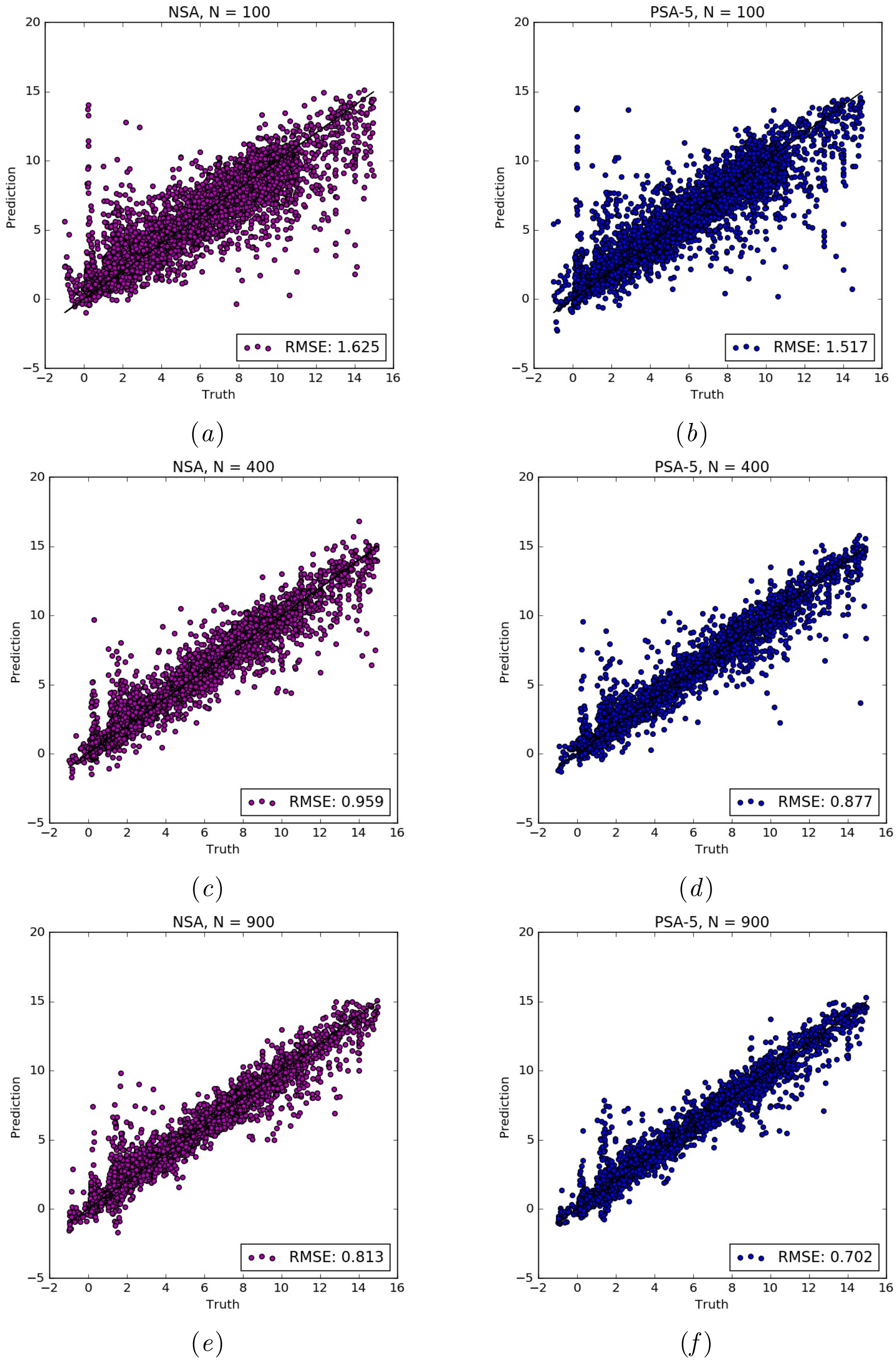
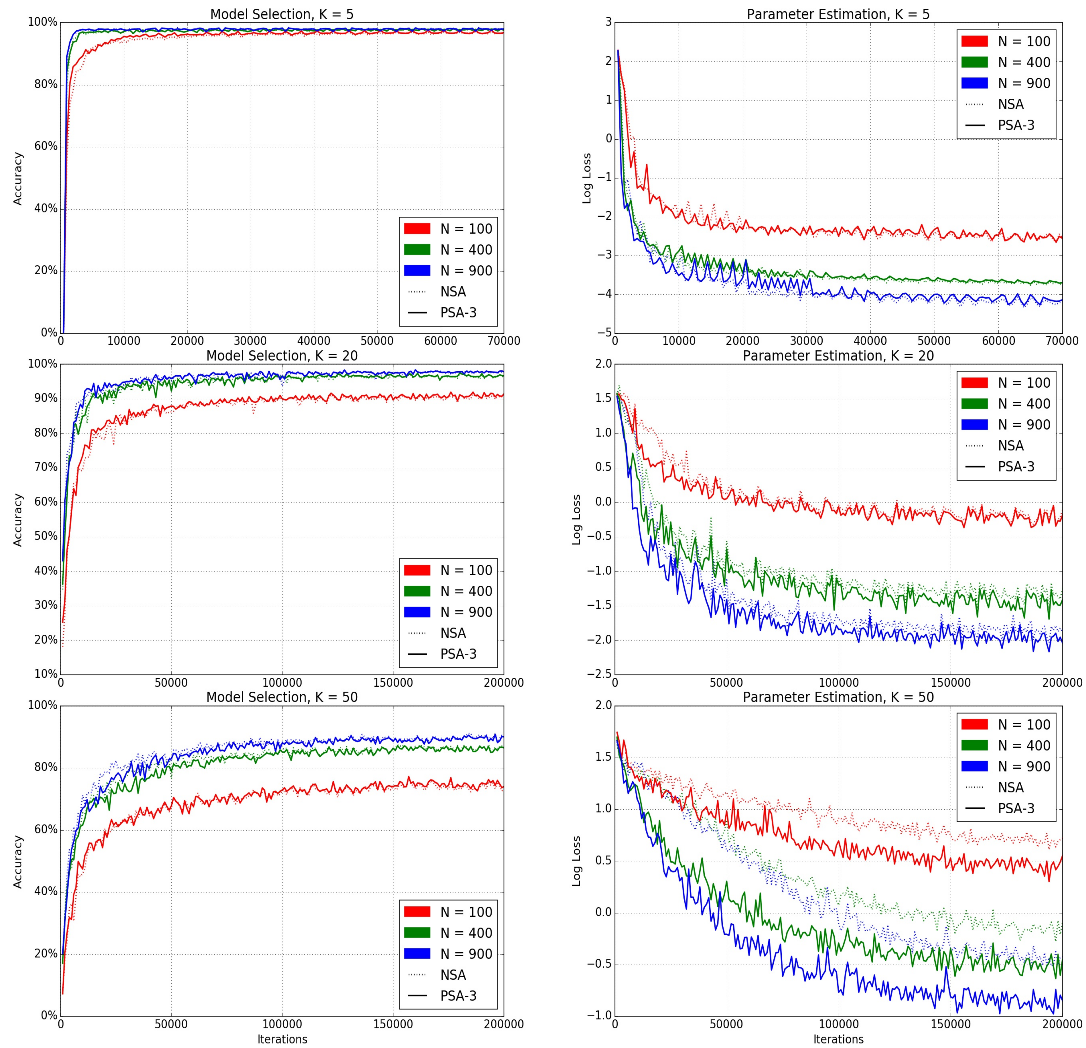
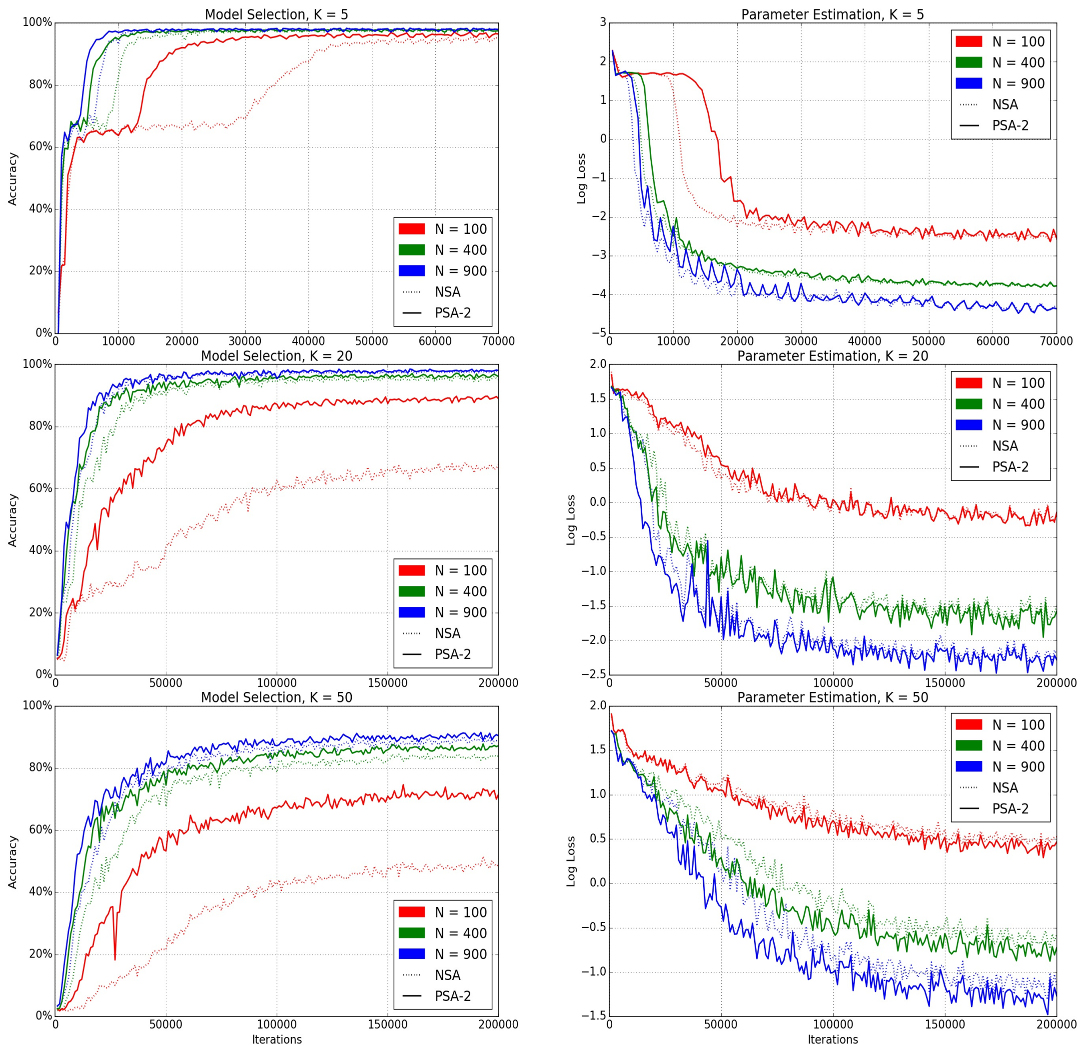


Figure 1: Parameter estimation results on the test dataset with $K = 50$. The left panel is the results estimated by large CNN and NSA parameter estimator, while the right panel is estimated by large CNN and PSA-5 parameter estimator. The x-axis in each plot is the ground truth for the parameters and the y-axis is the estimation.



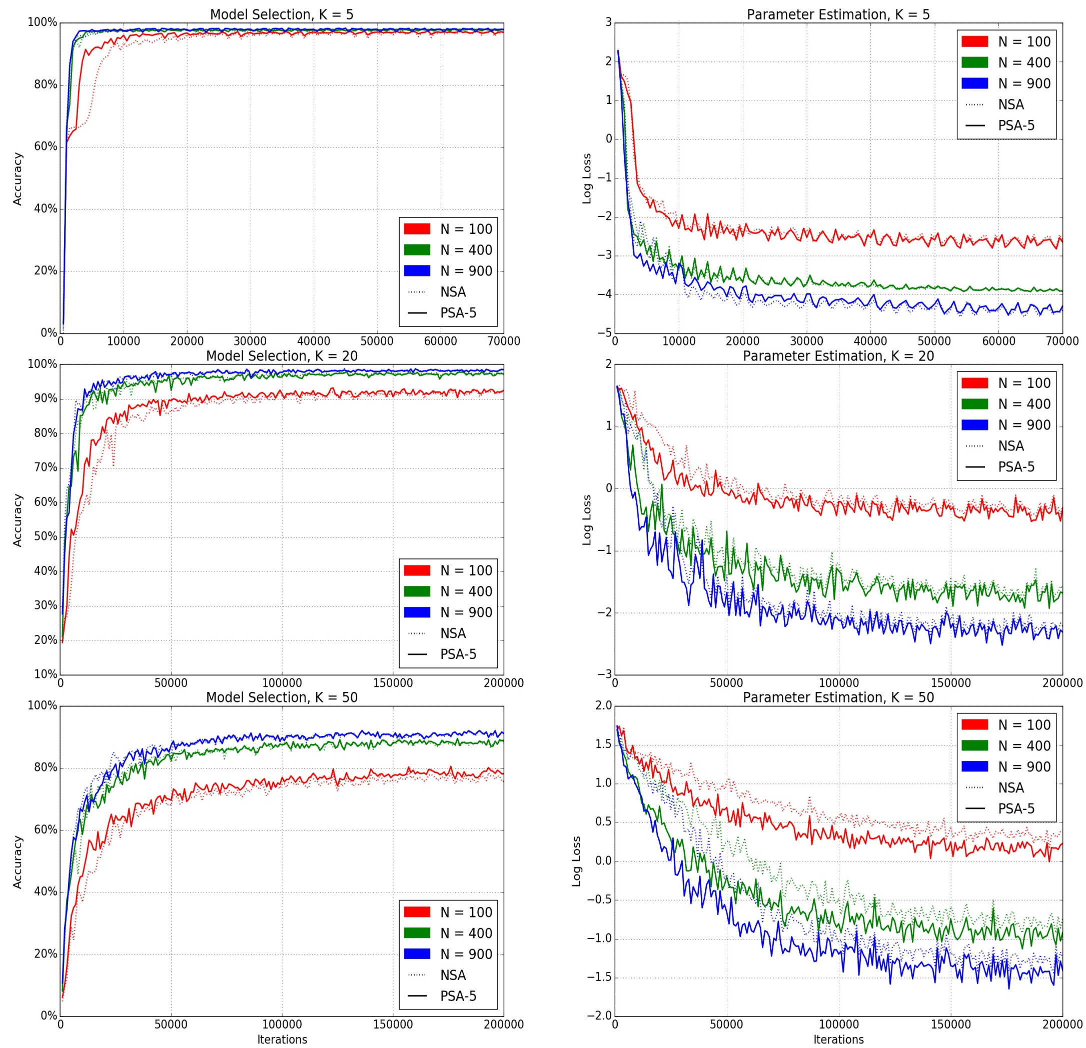
(a) Model Selection $\qquad$ (b) Parameter Estimation

Figure 2: Comparison between NSA and PSA-3: learning curves of selection accuracy and estimation Huber loss for different samples sizes and different number of candidate models. This is based on small CNN architecture. The left panel plots the selection accuracy of the model selector evaluated on the validation dataset against the number of iterations during the training process, whereas the right panel plots the log Huber loss of the parameter estimator on the validation dataset against the number of iterations during the training process. Solid curves are for the PSA-3, and dotted curves are for the NSA. Different colors denote for different sample sizes.



(a) Model Selection (b) Parameter Estimation

Figure 3: Comparison between NSA and PSA-2: learning curves of selection accuracy and estimation Huber loss for different samples sizes and different number of candidate models. This is based on medium CNN architecture. The left panel plots the selection accuracy of the model selector evaluated on the validation dataset against the number of iterations during the training process, whereas the right panel plots the log Huber loss of the parameter estimator on the validation dataset against the number of iterations during the training process. Solid curves are for the PSA-2, and dotted curves are for the NSA. Different colors denote for different sample sizes.



(*a*) Model Selection (*b*) Parameter Estimation

Figure 4: Comparison between NSA and PSA-5: learning curves of selection accuracy and estimation Huber loss for different samples sizes and different number of candidate models. This is based on large CNN architecture. The left panel plots the selection accuracy of the model selector evaluated on the validation dataset against the number of iterations during the training process, whereas the right panel plots the log Huber loss of the parameter estimator on the validation dataset against the number of iterations during the training process. Solid curves are for the PSA-5, and dotted curves are for the NSA. Different colors denote for different sample sizes.

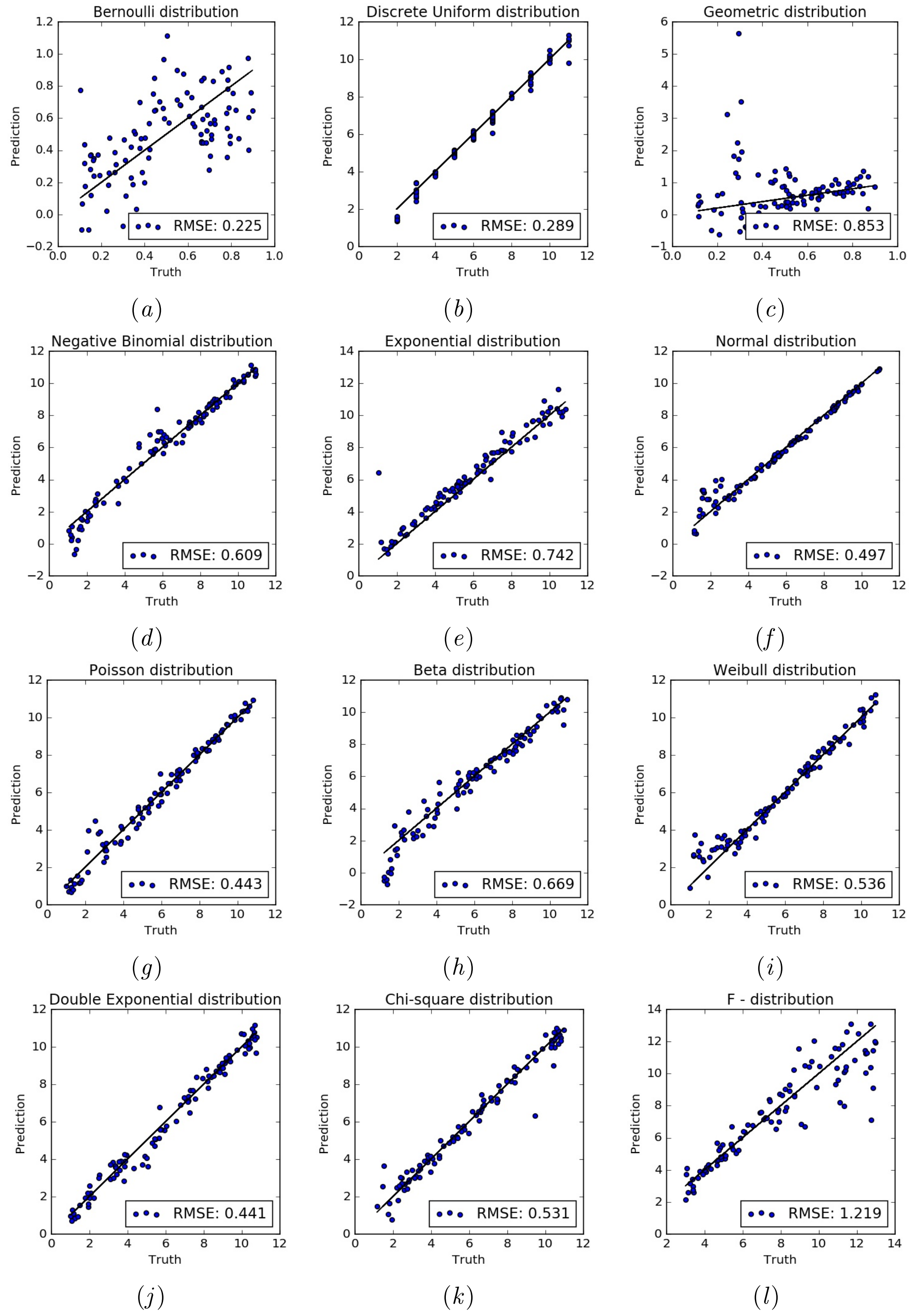


Figure 5: Distribution wise performance of PSA-5 neural parameter estimator under large CNN on the test dataset with $N = 900$. RMSE is reported for each distribution. Part 1.

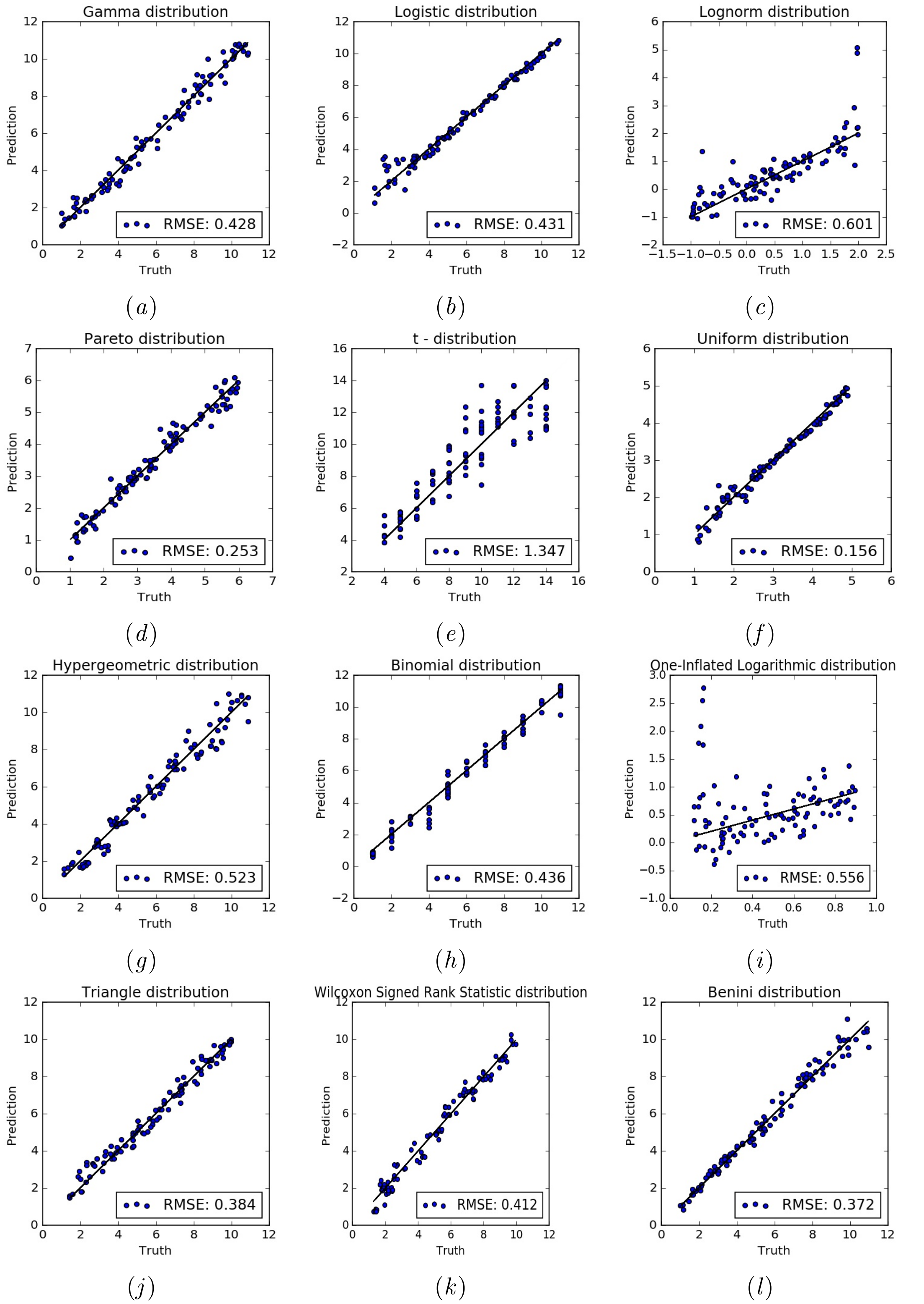


Figure 6: Distribution wise performance of PSA-5 neural parameter estimator under large CNN on the test dataset with $N = 900$. RMSE is reported for each distribution. Part 2.

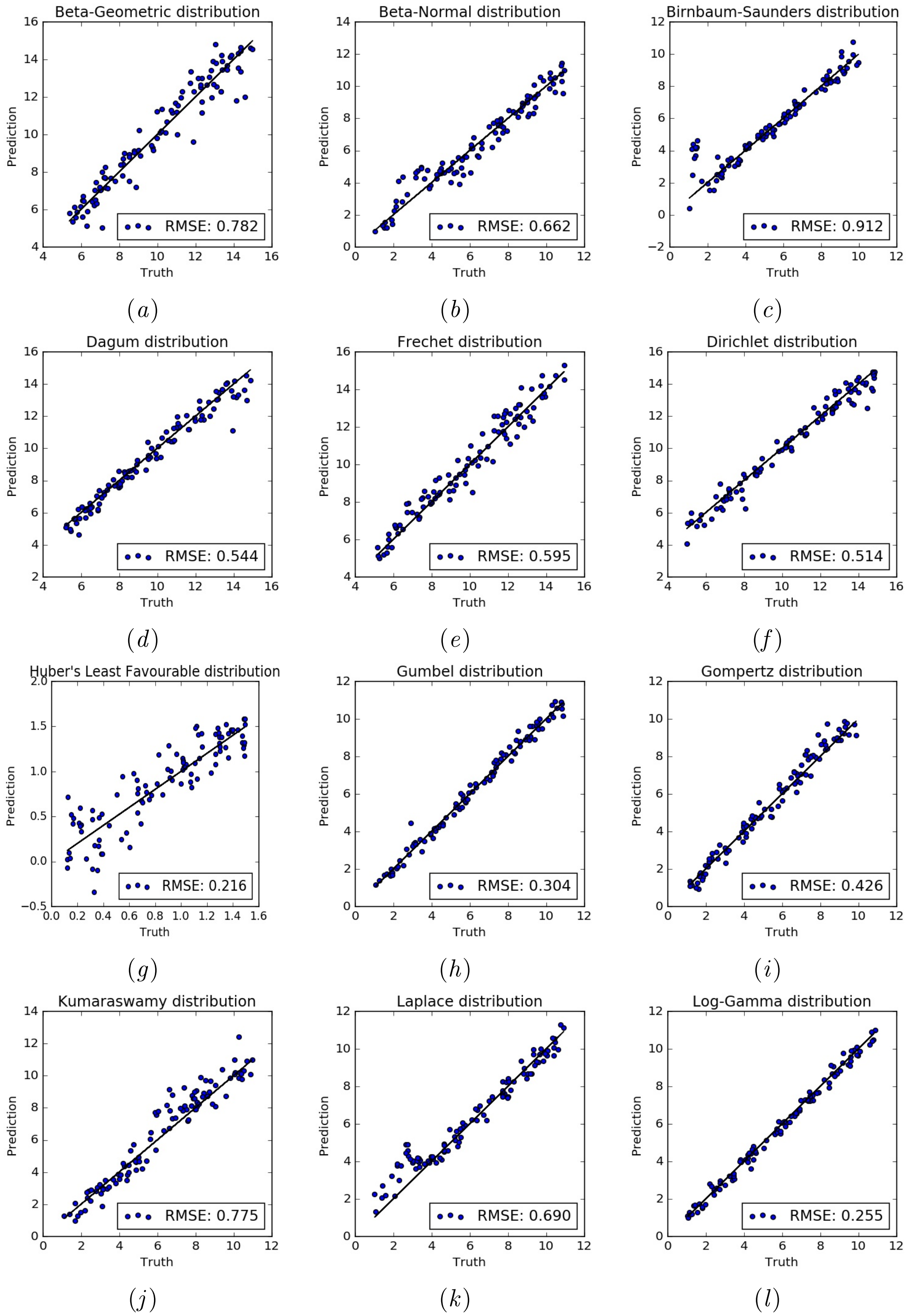


Figure 7: Distribution wise performance of PSA-5 neural parameter estimator under large CNN on the test dataset with $N = 900$. RMSE is reported for each distribution. Part 3.

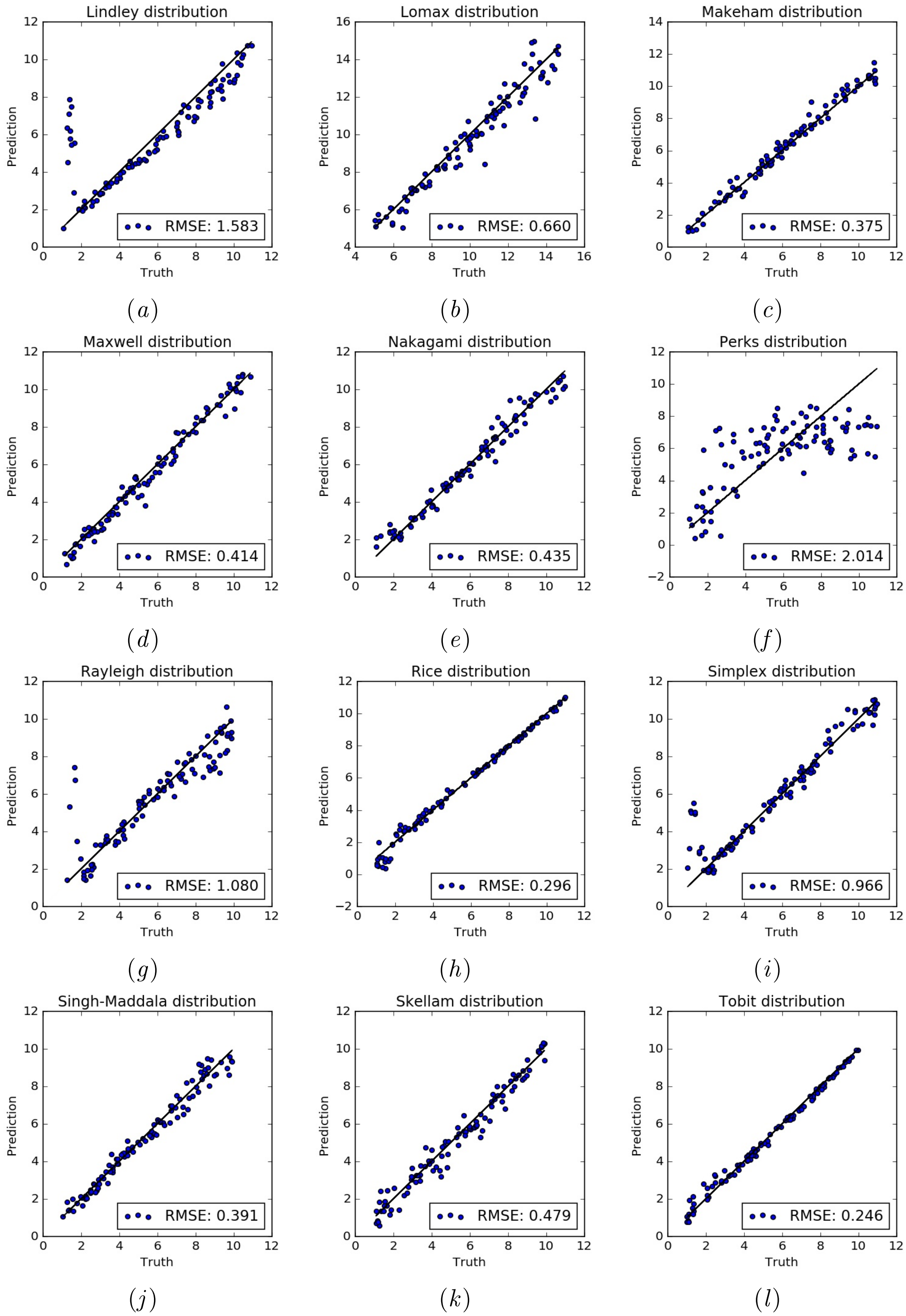
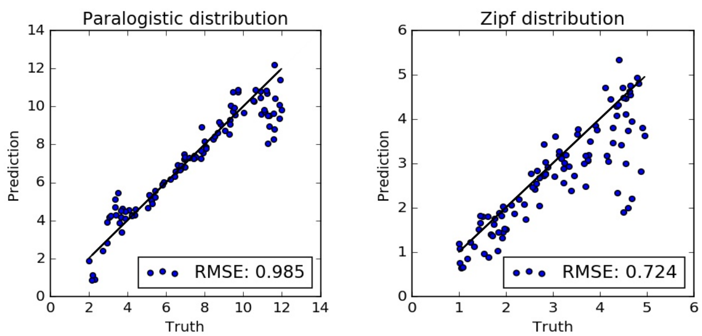


Figure 8: Distribution wise performance of PSA-5 neural parameter estimator under large CNN on the test dataset with $N = 900$. RMSE is reported for each distribution. Part 4.



(a) (b)

Figure 9: Distribution wise performance of PSA-5 neural parameter estimator under large CNN on the test dataset with $N = 900$. RMSE is reported for each distribution. Part 5.